# Predictive Analytics Using R and Python

- MUQUAYYAR AHMED
DATA SCIENTIST

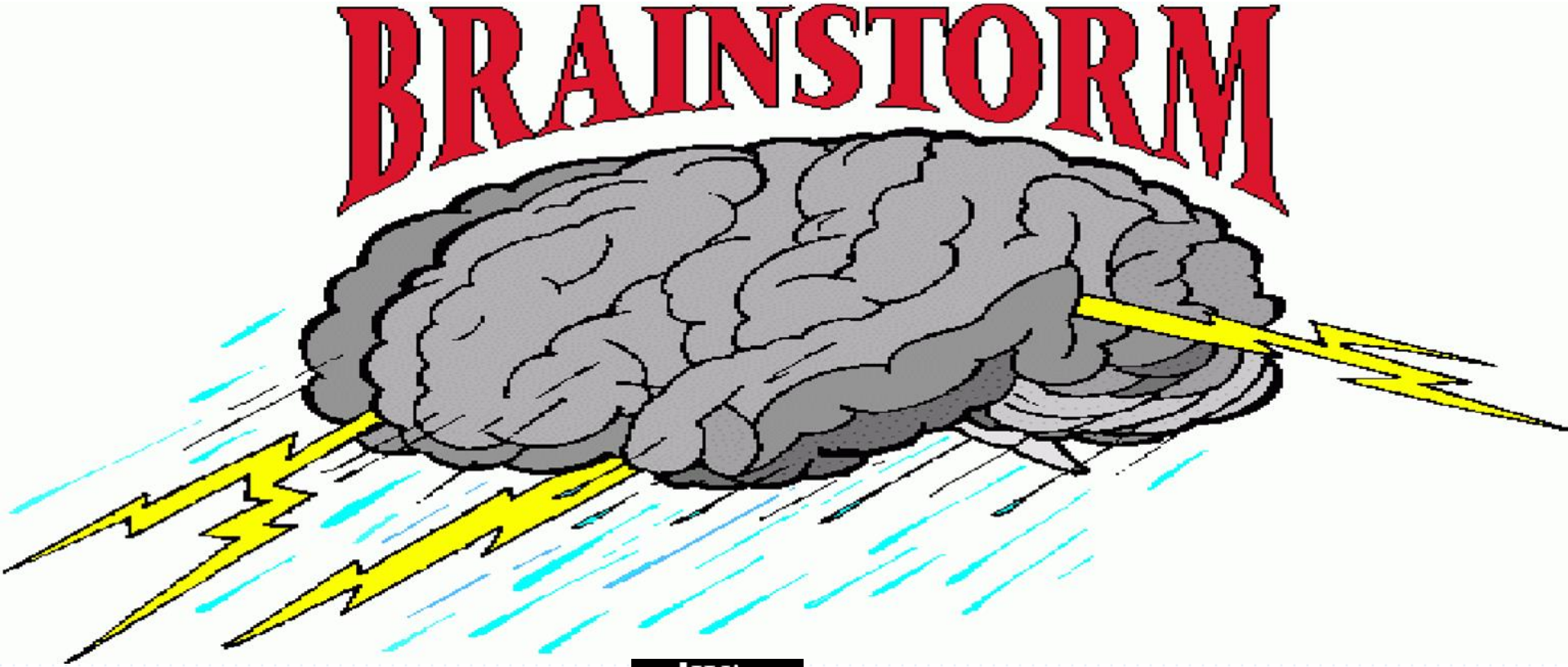
# We learnt!!

1. Introduction to Predictive Analytics
   - Skills Needed
   - What is Data Analytics
   - Applications
   - Type of Problems
   - CRISP-DM Process
2. Introduction to R
3. Introduction to Python

# Today's Agenda

- Exploratory Data Analysis

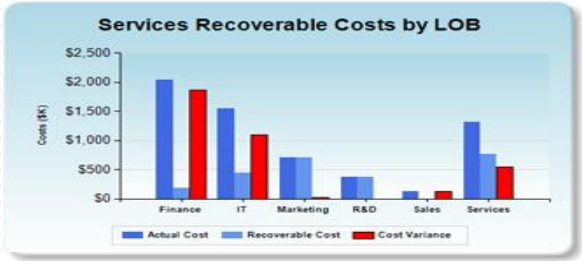
# Brain Storming Session



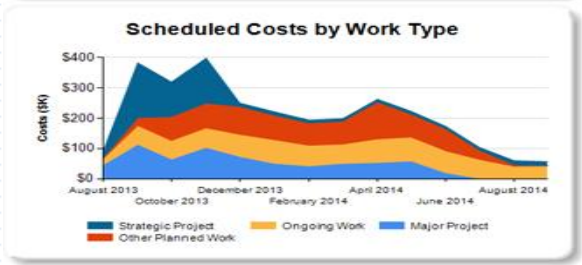
# Exploratory Data Analysis

## Objective of data exploration

- Mathematical equations
- Distribution of a variable
- Relation between attributes
- Comparison
- Trend as a function of time








# Data types

## Two types of variables

- Categorical
  - Ordinal. Ex: Low, medium, high
  - Nominal. Ex: Tiger, Lion, Elephant
- Continuous. Ex: 4, 5, 3.2

## Different types of data

- Factor
- Character
- Numeric

# Changing the type

**Neural networks**

- All attributes must be numeric

**Naive Bayes**

- All attributes must be categorical

# Changing numeric to categorical

Equal frequency (Number of samples in each bin is equal)

- Let us say, we have the data for price: 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- If we partition into equal-frequency (equi-depth) bins we get the following three bins
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34

Equal width (Interval is same (good for uniform distributions))

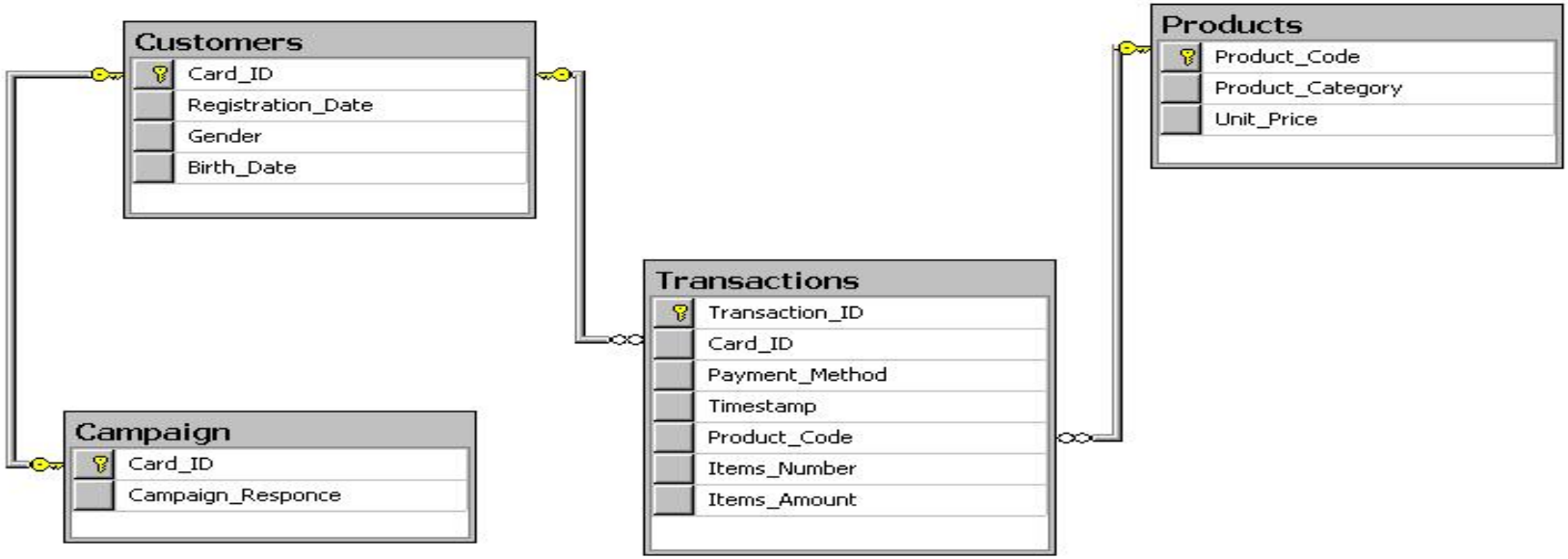
- If we partition into equal width then we get
  - Bin 1: 4, 8, 9
  - Bin 2: 15, 21, 21, 24
  - Bin 3: 25, 26, 28, 29, 34

# Changing categorical to numeric

- Five codes for marital status (“single,” “divorced,” “married,” “widowed,” and “unknown”) would be mapped to –1.0, –0.5, 0.0, +0.5, +1.0, respectively

# Merging and sorting of data

Entity relationship diagram (ER diagram)



# R Session

- Data Frame
- vectors
- Matrices
- Lists
- Data types
- Sorting
- Merging
- Cbind and rbind
- Data summary
- Activity

edWisor

# Learn R

- http://www.r-bloggers.com/
- http://www.statmethods.net/
- http://rfunction.com/